

A Evaluation And Experiments

This section presents a comprehensive evaluation based on MeetBench of the MeetMaster system using the MeetALL dataset dataset. We compare MeetMaster against several baseline models, analyze the performance using multiple metrics, and conduct ablation studies to understand the contribution of each module. Additionally, we provide detailed implementation insights to ensure reproducibility. The experiments are conducted on the train_S consisting of 53 conference recordings, each of which was approximately 30 minutes long, as described in Section 3.1.

Table 6: Ablation Study Results on Five Dimensions of MeetBench

Model	Factual	User Needs	Conciseness	Structure	Completeness	Overall Score
Only Talker	5.96	5.25	6.27	6.03	5.00	5.38
Only Planner	7.99	6.99	8.32	7.76	6.39	7.05
MeetMaster	7.50	6.57	6.76	7.33	6.36	6.59

A.1 Baseline Methods

To benchmark MeetMaster’s performance, we compared it against several baseline models: LLAMA-7B and LLAMA-13B [26], Qwen2-Audio (a multimodal audio-focused model), Qwen2.5-7B (7B parameters with enhanced reasoning), ChatGLM3-6B (6B parameters optimized for conversations), DeepSeek-R1’s 7B, Qwen-Agent and Phi-1 & Phi-1.5 [17]. These diverse models provide a comprehensive evaluation framework for MeetMaster’s capabilities.

A.2 Evaluation Metrics

As described in Section 4.1 and 4.2, we employ CompassJuder and Prometheus frameworks for comprehensive evaluation, along with latency measurements to assess system performance in real-time scenarios.

A.3 Implementation Details

The MeetMaster system was implemented using Python 3.11.4. All experiments were conducted on 2×NVIDIA RTX 3090 GPUs with PyTorch 2.1.0. In terms of hardware configuration, one GPU was dedicated to the Talker module while the other handled the Planner module, enabling parallel processing and efficient resource utilization.

A.4 Experimental Results

A.4.1 Quality and Accuracy Measurements. Tables 4 presents the evaluation results of MeetMaster and baseline models on MeetALL dataset using both CompassJuder and Prometheus frameworks. MeetMaster achieves competitive performance across all evaluation dimensions, scoring 6.59 in CompassJuder’s overall assessment and 3.50 in Prometheus evaluation. MeetMaster’s scores demonstrate its effectiveness in balancing response quality with real-time meeting assistance requirements.

Table 7: Latency Measurements for MeetMaster

Component	Latency (ms)
STT Module (per Token)	53
Talker Latency (first Token)	210
Talker Latency (each Token)	31
Planner Latency (first Token)	520
Planner Latency (each Token)	310

Table 8: Ablation Study Results on Overall Score

Model	Prometheus Score
MeetMaster	6.59
Talker	5.14
Planner	6.98

A.4.2 Latency Measurements. Latency is a critical factor in real-time meeting assistance systems. Table 7 summarizes the latency results for MeetMaster, highlighting its efficiency in handling both simple and complex queries.

The STT module processes simple queries in approximately 550 milliseconds and complex queries in about 1600 milliseconds. Whisper streaming maintains a consistent token processing time of 50 milliseconds, ensuring rapid transcription. The Talker module responds to the first token within 200 milliseconds and generates each subsequent token in 30 milliseconds. The Planner module processes each token with a latency of 300 milliseconds, balancing speed and depth in response generation.

Our MeetMaster system represents a strategic trade-off between response time and processing complexity. When using only our Planner, the system requires 15-20 seconds for both simple and complex queries due to its comprehensive reasoning process. By incorporating our Talker, MeetMaster achieves end-to-end response times of just 900ms for straightforward queries while maintaining the 15-20 second processing time for complex tasks that require deeper reasoning. This dual-process approach effectively reduces response times by up to 1500ms (from 15 seconds to 900ms) for routine queries compared to a single-model system, while preserving the thorough analysis capabilities necessary for complex scenarios.

A.4.3 Ablation study of individual contribution. To assess the individual contributions of the Talker and Planner modules, we conducted ablation studies by evaluating the system with only one of the modules active. The results are presented in Table 8 and in Table 6. Compared to Only Talker, MeetMaster delivers 22.5% better outcomes on overall tasks and compared to Only Planner, MeetMaster cuts the fastest response times for routine queries by 1500 ms while processing simple tasks.

The ablation results indicate that the Planner module alone significantly outperforms the Talker module alone in all evaluation dimensions. However, the combined MeetMaster system leverages the strengths of both modules, achieving a balanced and superior overall performance compared to using either module individually.

Combining the CompassJugder and Prometheus scores with latency measurements, MeetMaster demonstrates robust performance across multiple facets of meeting assistance. MeetMaster outperforms all baseline models in factual accuracy, user need fulfillment, and structural clarity, while maintaining competitive scores in conciseness. The Prometheus scores further validate MeetMaster’s superior overall performance in real-world meeting scenarios.

Latency measurements reveal that MeetMaster maintains low processing times for simple queries, ensuring real-time responsiveness. The efficient Whisper streaming and optimized Talker and Planner modules contribute to the system’s ability to deliver timely and accurate responses, enhancing the overall meeting experience.

B Implementation details of MeetMaster and benchmark

B.1 Speech-to-Text and Whisper-Based Tooling

A cornerstone of any effective AI meeting assistant is its ability to accurately convert live speech into textual form. Recently, **Whisper** [22] has emerged as a powerful open-source automatic speech recognition (ASR) model capable of streaming transcription with low latency, making it a strong candidate for real-time meeting scenarios. Whisper’s encoder-decoder architecture utilizes a multi-layer Transformer to capture both local acoustic patterns and longer-range language dependencies, enabling robust performance even with overlapping speech segments.

Several studies have extended the Whisper framework with domain-specific fine-tuning or integrated it into larger toolchains for conversation management. For instance, [19] demonstrated that streaming Whisper modules could seamlessly interface with knowledge-graph queries in corporate meeting systems. In our MeetMaster pipeline, Whisper provides rapid voice-to-text conversion for the *Planner* agent. Moreover, we incorporate streaming decoding to minimize latency when handling short, routine queries. By leveraging Whisper’s open-source codebase, developers can fine-tune or adapt the ASR stage for specialized vocabularies, enhancing domain coverage and transcription accuracy in dynamic meeting contexts.

B.2 AISHELL Dataset and Extensions.

To develop robust Mandarin TTS or automatic speech recognition (ASR) systems, the AISHELL-1 corpus [3] has become a standard benchmark. It contains 400 speakers recorded with consistent audio quality, accompanied by accurate manual transcriptions under the Apache License v2.0. While originally proposed for ASR research, AISHELL has proven valuable for broader speech technologies, including speech synthesis and speaker adaptation. Building on this foundation, various extended datasets have emerged to address more specialized needs, such as accented speech, spontaneous dialogue, or multi-speaker interactions [9, 12, 24]. In our work, we introduce *MeetALL dataset*—an enhanced version incorporating both simple and complex meeting scenarios, agent wake-word triggers, and voice-cloned system responses. This extension provides a more diverse set of contextual utterances, broadening the use of AISHELL beyond standard ASR tasks to complex meeting assistance.

Meeting STT Data in TextGrid and JSONL Format: The speech-to-text (STT) data for each meeting is stored in TextGrid format and JSONL Format. TextGrid is a widely used file format in meeting audio datasets, where the transcribed text is aligned with the corresponding audio timestamps. This allows for accurate time-alignment and retrieval of the meeting content, ensuring that the generated requests and responses are contextually relevant and time-sensitive.

B.3 Fast and Slow Thinking Agent Architectures.

Inspired by Kahneman’s *Thinking, Fast and Slow* [6], recent research has explored dual-process frameworks for conversational AI. These systems encapsulate rapid, intuitive responses (System1) and slower, more deliberative reasoning (System2). For instance, propose a *Talker-Reasoner* setup to manage conversational flow and background planning. This architecture is particularly advantageous in interactive settings, where straightforward requests can be handled quickly to enhance user experience, while complex questions invoke a deeper chain of reasoning and tool usage. Related ideas appear in works that attempt to integrate retrieval-augmented generation (RAG) for knowledge-intensive tasks [21], ensuring that the more deliberative agent has access to relevant documents or APIs. Our *MeetMaster* system generalizes these paradigms, coupling a fast-thinking “Talker” for simple queries with a slow-thinking “Planner” capable of multi-step problem solving in meeting scenarios.

B.4 Audio-based Keyword Spotting

In text-only recognition systems, environmental noise or ASR errors often lead to missed or false detections of the agent-triggered keyword. To enhance wake-up robustness, *MeetMaster* supplements parallel text detection with audio signal-based keyword detection to capture the specific acoustic pattern. This module adopts end-to-end miniaturized neural networks methodology from WeKws[28], enabling real-time keyword discrimination on streaming input and complementary fusion with text-triggers. Through this parallel detection system, *MeetMaster* achieves rapid, accurate instruction recognition in noisy meeting environments or multi-channel interaction scenarios, enabling reliable activation of fast and slow-thinking agents and significantly improving the overall meeting assistant system’s interaction success rate.

B.5 Automated Evaluation of LLM Outputs.

The rapid proliferation of large language models has spurred the development of automated evaluators to assess text quality, correctness, and alignment [11]. Two notable open-source evaluators are *CompassJugder*[4] and *Prometheus*[15, 16]. *CompassJugder* aims to unify single-model scoring (rating an output against a rubric) and pairwise comparison (choosing the better of two outputs). It produces dimension-specific feedback (e.g., Factual Accuracy, User Needs, Conciseness, Structure, Completeness, Overall Score) and is thus well-suited for multi-criteria evaluation. *Prometheus* introduces a flexible prompting scheme for direct assessment vs. pairwise ranking, and it can incorporate custom evaluation standards beyond simple helpfulness. Our work adapts these frameworks to the domain of meeting assistance by introducing specialized prompts that account for meeting context, timeliness, and complexity handling. Through this dual-evaluator strategy, we obtain a holistic view of a system’s performance in both routine and deliberative tasks.

C Experimental Results and System Evaluation

C.1 Test Meeting Example

The dialogue below shows how *MeetMaster* routes three real meeting questions through the four-axis classifier → Talker / Planner pipeline and how the Planner’s three-hop loop (*Analyse* → *Reasoning* → *Compose & Act*) works in practice.

For brevity, the JSON exchanged between hops is partially truncated. To facilitate recognition, parts of the example is color-coded: **violet** is hop a (*Analyze*), **teal** is hop b (*Planning*), and **olive** is hop c (*Compose & Act*) and final output.

Q1 (Complex, routed to Planner)

USER: Hello, JiaoJiao! Based on our previous discussions, how can we strategically improve the organisation of the tournament to ensure a seamless and engaging experience for players, sponsors and audiences?

Planner output {"category": {"CL": "High", "CD": "Cross-Meeting", "DK": "Specialist", "TK": "Organise"}, "level": "Complex", "keywords": ["tournament", "audience engagement", "sponsor"]}

PLANNER – Hop a) Analyze (above)

PLANNER – Hop b) Planning

[{"tool": "RagRetrieve", "source": "KB-42", "snippet": "... points-based | qualification improves fairness...", ...}]

PLANNER – Hop c) Compose & Act → final answer

- **Tournament structure** – points-based qualifiers → top-16 knock-out → best-of-five final [1]
- **Audience engagement** – 1-2 h match blocks; weekend finals; bilingual live-stream [2]
- **Sponsor value** – branded replay clips, VIP lounge, on-site product booths [3]

Action: SCHEDULEMEETING with Ops & Marketing at 14 May 2025 09:00 JST

Q2 (Simple, routed to Talker)

USER: Hello, JiaoJiao! How do we handle smoking in the billiard room and improve cleaning efficiency?

Planner output **T!**

TALKER : “Please designate an outdoor smoking spot and schedule hourly mini-clean-ups; staff can politely guide smokers to the area.”

Q3 (Complex summary request, routed to Planner)

USER: Hello, JiaoJiao! Could you summarize all decisions and action items about value-added services and sponsorship strategies discussed today, including timelines and owners?

Planner output {"category":{"CL":"High", "CD":"Cross-Meeting", "DK":"Specialist", "TK":"Organise"},
"level":"Complex", "keywords":["value-added service", "sponsorship", "timeline"]}

PLANNER – Hop a) Analyse (above)

PLANNER – Hop b) Planning calls MEETSUM on the entire transcript and RAGRETRIEVE on company KB (6 snippets total).

PLANNER – Hop c) Compose & Act → final answer (excerpt) [noitemsep,topsep=0pt]

- **Value-Added Services**

- Launch 1 Aug 2025 • Owner Marketing • Target 30 % uptake, +15 % rev [1]

- **Sponsorship**

- Pitch deck ready 15 Jun 2025

- Owner Events

- Goal 5 new sponsors [2]

Integrated timeline – bi-weekly progress stand-ups;

final review Dec 2025 [3]

Action: GENERATESLIDE (“Sponsorship Roadmap Q3-Q4 2025”, 4 slides)

This example demonstrates that Talker answers simple questions in low context in subsecond latency, while high-stakes cross-meeting queries trigger Planner’s evidence-based three-hop reasoning: delivering strategic results with concrete actions while keeping the overall flow of the meetings uninterrupted.

D Query Routing Mechanism

MeetMaster’s decision-making for meeting queries is governed by Algorithm 1.

Algorithm 1 Query Routing with Fine-tuned Planner Classifier

```

1: procedure HANDLEMEETINGQUERY(userQuery)
2:   Initialise planner and talker agents
3:   (cat, level) ← PLANNER.CLASSIFY(userQuery)
4:   if level = SIMPLE and cat ∈ {CLLow, CDNone, DKGeneral, TKRecord} then
5:     return TALKER.QUICKRESPONSE(userQuery)
6:   else
7:     plan ← PLANNER.PLANCARD(userQuery)
8:     answer ← PLANNER.REASON(plan)
9:     return FORMATRESPONSE(answer)
10:  end if
11: end procedure

```

► single forward pass

► tool chain, evidence, stop

► three-hop loop

D.1 Real-time Delay Experiment

Here is another example demonstrating the system’s response time for processing questions. As shown in the following output, when processing simple question by talker, end-to-end response of MeetMaster is $1737102519.7489702 - 1737102518.779651 = 969$ ms:

[Agent] Final received question: Hello, teach me how to increase the sales of inventory clothing through online channels and predict market trends.
Time when the question audio was sent to Planner: 1737102518.779651
Warning: do_sample is set to False. However, top_p is set to 0.9 this flag is only used in sample-based generation modes. You should set do_sample=True or unset top_p.
planner new_token time: 1737102519.4024565
First token received: 1, time: 1737102519.7965689
Judging it as a short question, planner output ends, time: time.time()
planner_output_list: ['1']
Selected agent: talker
Inputting question audio to talker
talker first token time: 1737102519.7489702
talker output:
system
You are a helpful assistant.
user
Answer the voice question in less than 100 words. Audio 1:
assistant
Hello, you can increase inventory clothing sales and predict market trends by:
Using online channels for promotions like discounts, buy one get one free, etc.
Promoting through social media to increase brand exposure.
Regularly updating product styles to meet market demand.
Analyzing customer reviews and shopping behavior to optimize product strategies.
Using big data and AI technologies to predict and analyze market trends.
talker output ends
Processing file: /home/leon/agent/AISHELL_dataset/insert_train_S/20200807_S_R001S04C01_agent_added/out_001-F_1.wav
Question audio length: 17.2143125
[Agent] Final received question: Hello, Jiao Jiao, based on what we discussed earlier about improving brand image and attracting consumers, what do you think of using a combination of amateur models and professional models for new product showcases and fashion events? Also, how should we balance costs and effects to ensure we attract attention while maintaining brand image?
Time question audio was sent to Planner: 1737102526.4261591
planner new_token time: 1737102527.0768712
planner output starts:
Action: Information retrieval RAG
Action input: Keywords: amateur models, professional models, new product showcases, cost-effectiveness, balancing cost and effect
Extracted keywords: ['amateur models', 'professional models', 'new product showcases', 'cost-effectiveness', 'balancing cost and effect']
Found context fragments: 2
planner_output_list: ['0', '1. ', 'Action: Information retrieval RAG', '2. ', 'Action input: Keywords: amateur models, professional models, new product showcases, cost-effectiveness, balancing cost and effect', 'RAG used context: ...']
Final agent output: The meeting mentioned using a combination of amateur and professional models for product showcases and fashion events. Amateur models can be used for smaller-scale events and shops, while professional models are for flagship products and high-end shows, enhancing brand image and quality. Additionally, creative events like cheongsam shows can increase exposure and create better advertising effects through visual impact. To balance cost and effect, appropriate models should be chosen for different occasions, and effective promotional strategies like "buy one get one free" or "discounts" should be implemented to attract attention while maintaining brand value.
End of planner output.

D.2 Dataset creation details and Dataset Format

The dataset used in our experiments follows a structure based on the AISHELL dataset, formatted in a standard text grid format. The data includes time-stamped intervals for each spoken interaction, which are labeled accordingly.

Example of a data segment:

File type = "ooTextFile"

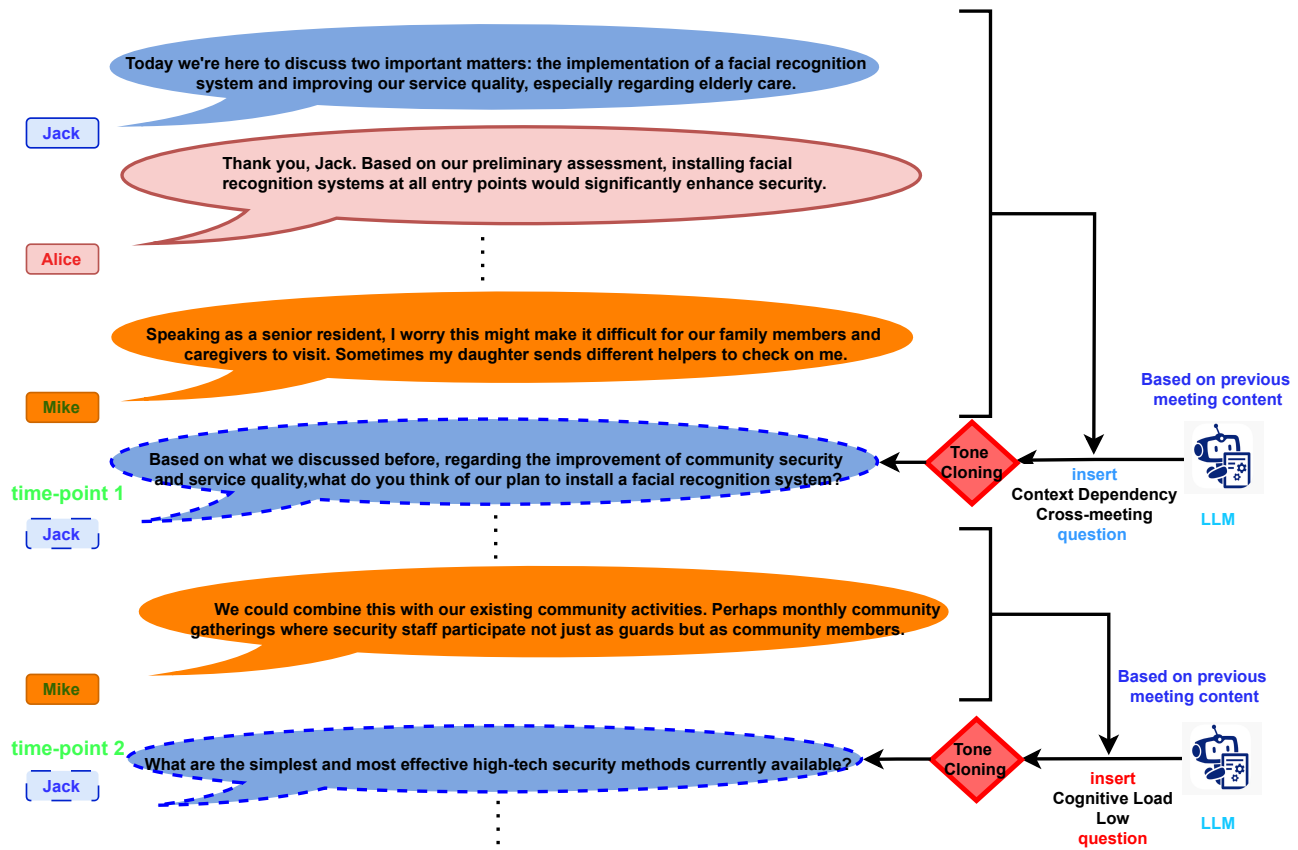


Figure 4: Illustration of MeetALL dataset construction methodology, demonstrating the injection of AI agent interactions into meeting dialogues. The process incorporates tone cloning technology and at each time-point we generate one type query with all complexities (choosing type from Cognitive Load, Context Dependency, Domain Knowledge and Task-Execution Effort). This dataset construction approach ensures natural conversation flow while maintaining contextual relevance and query complexity differentiation.

```
Object class = "TextGrid"
xmin = 0.0
xmax = 1949.076
tiers? <exists>
size = 4
item []:
  item [1]:
    class = "IntervalTier"
    name = "006-M"
    xmin = 0.0
    xmax = 1949.076
    intervals: size = 173
    intervals [1]:
      xmin = 0.0
      xmax = 18.305
      text = ""
    intervals [2]:
      xmin = 18.305
      xmax = 22.08
      text = "002 I'm the investment teacher"
```



```

intervals [3]:
    xmin = 22.08
    xmax = 31.23633
    text = ""
intervals [4]:
    xmin = 31.23633
    xmax = 38.11133
    text = "Okay, good. Today I called everyone here to <sil> discuss together..."
...
intervals [79]:
    xmin = 436.9628
    xmax = 497.63514
    text = ""
intervals [80]:
    xmin = 497.63514
    xmax = 507.80014
    text = "Oh then you're a second-generation rich! If that's the case, <sil>
your family has sufficient financial support then you can..."
intervals [81]:
    xmin = 507.80014
    xmax = 507.80014
    text = "Hello Jiaojiao, how should college students balance funding, team and market
demand when starting a business?"
...
intervals [208]:
    xmin = 992.50424
    xmax = 1104.67465
    text = ""
intervals [209]:
    xmin = 1104.67465
    xmax = 1104.67465
    text = "Hello Jiaojiao, based on our previous discussion about challenges in breeding
industry, especially pandemic impacts, how should we implement scientific management to
reduce losses while maintaining pork quality? Also, how to formulate pricing strategies
for different consumer classes?"
intervals [210]:
    xmin = 1104.67465
    xmax = 1106.05717
    text = "That's a good idea,"
...

```

D.3 Groud Truth Examplexs and Reason why it belongs to such complexity level

Low complexity example:
Question: "What type of tea does P20 offer to P18?"
Ground truth answer: "P20 offers green tea to P18."
The complexity level: "Low complexity"
Explanation of why it belongs to that complexity level: "This question is classified as low complexity because it requires simple recall of an explicitly stated fact from the transcript. The answer can be directly retrieved from P20's statement without requiring any synthesis, comparison, analysis, or inference. It only asks for a specific type of tea (green tea) that was clearly mentioned in the meeting."

Medium complexity example:
Question: "What are the different opinions about kombucha expressed in the meeting?"
Ground truth answer: "P20 mentions that they have tried kombucha but are not a big fan due to the fermented flavor, while P17 inquires about the sourness or flavor, suggesting a curiosity or openness to it."
The complexity level: "Medium complexity"
Explanation of why it belongs to that complexity level: "This question is classified as medium complexity because it requires synthesizing information from different parts of the discussion. The answer involves gathering P20's opinion about kombucha and P17's inquiry about its taste, which are mentioned separately in the transcript. This involves connecting related pieces of information to understand the different opinions expressed."

High complexity example:
Question: "What might be the reasons for P20's preference for yogurt over kombucha despite both being fermented products?"
Ground truth answer: "P20 prefers yogurt over kombucha because they find the fermented flavor in drinks unappealing, whereas they enjoy yogurt and similar products like 'pathal' due to their sweetness. This suggests that P20's preference might be influenced by the sweetness balancing the fermented taste, which is more pronounced in yogurt than in kombucha."
The complexity level: "High complexity"
Explanation of why it belongs to that complexity level: "This question is classified as high complexity because it requires analysis, evaluation, and inference beyond what is explicitly stated in the transcript. Answering requires understanding P20's preferences and inferring that the sweetness in yogurt might mitigate the fermented flavor that they dislike in kombucha. This involves evaluating the reasons behind P20's preferences and making inferences about taste profiles and personal preferences."

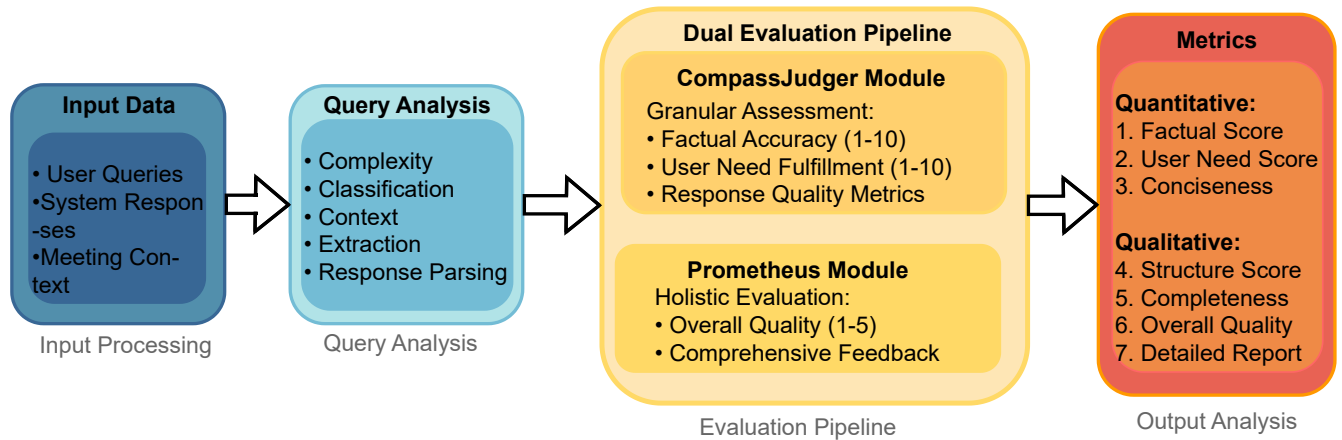


Figure 5: Schematic representation of MeetBench, illustrating the comprehensive evaluation pipeline. The framework integrates CompassJugder and Prometheus modules for granular assessment of agent responses across multiple performance dimensions.

D.4 Evaluation Process

Our evaluation process begins with query classification, where requests are categorized as different complexity to trigger appropriate evaluation criteria. Each response is then matched with its corresponding ground truth for accuracy assessment. The evaluation leverages two independent frameworks: CompassJugder for fine-grained dimensional analysis and Prometheus for holistic assessment with detailed feedback. This dual-framework approach enables robust evaluation of both response quality and agent behavior, providing comprehensive insights for continuous improvement of meeting assistance systems.

E CompassJugder Evaluation of AI Assistant's Responses

E.1 Query 1: How can innovation improve the competitiveness of the real estate and renovation industries?

Prompt: You are an assistant skilled in evaluating the quality of responses from meeting agents. Please evaluate the quality of an AI assistant's response to the user's question as an impartial judge. Since the answer involves role-playing, you need to assess it based on the following dimensions:

- (1) **Accuracy of Facts:** Is the information provided in the answer accurate and based on credible facts and data?
- (2) **User's Needs Fulfillment:** Does the answer meet the user's purpose and needs, and does it provide a comprehensive and appropriate response to the question?
- (3) **Conciseness:** Is the answer concise and clear, avoiding redundancy and repetition? A brief answer is important and will score higher.
- (4) **Structure:** Is the answer well-organized and easy to understand quickly, with a clear focus on the main points?
- (5) **Completeness:** Does the answer cover most of the key points related to the question, without omitting important information?

We will provide the user's question, a high-quality reference answer, and the AI assistant's answer for evaluation. If the reference answer states that relevant content is not mentioned but the assistant's answer includes relevant information, the assistant's answer should be rated higher, possibly achieving a perfect score of 10.

When you begin your evaluation, follow this process:

- (1) Compare the AI assistant's answer with the reference answer, highlighting any shortcomings in the assistant's response, and explain further. Keep in mind that the reference answer may not always be better than the assistant's answer. The reference answer's quality level ranges from 3 to 10, so judge accordingly.
 - (2) Evaluate the AI assistant's answer in different dimensions, giving a score between 1 and 10 for each.
 - (3) Finally, provide a comprehensive score based on the evaluations in each dimension.
 - (4) Your scoring must be as strict as possible, following these rules:
 - Significant factual errors or irrelevant/harmful content should result in a low score (1-2).
 - Answers with no serious errors but of low quality should receive a low score (3-4).
 - Answers that generally meet the requirements but are weaker in some dimensions can receive a middle score (5-7).
 - Answers that are similar to or slightly worse than the reference answer can receive a score of 8-9.
 - Answers that surpass the reference answer by being more concise and nearly perfect in all dimensions can receive a score of 10.
- Example: A reference answer can score 10 points.

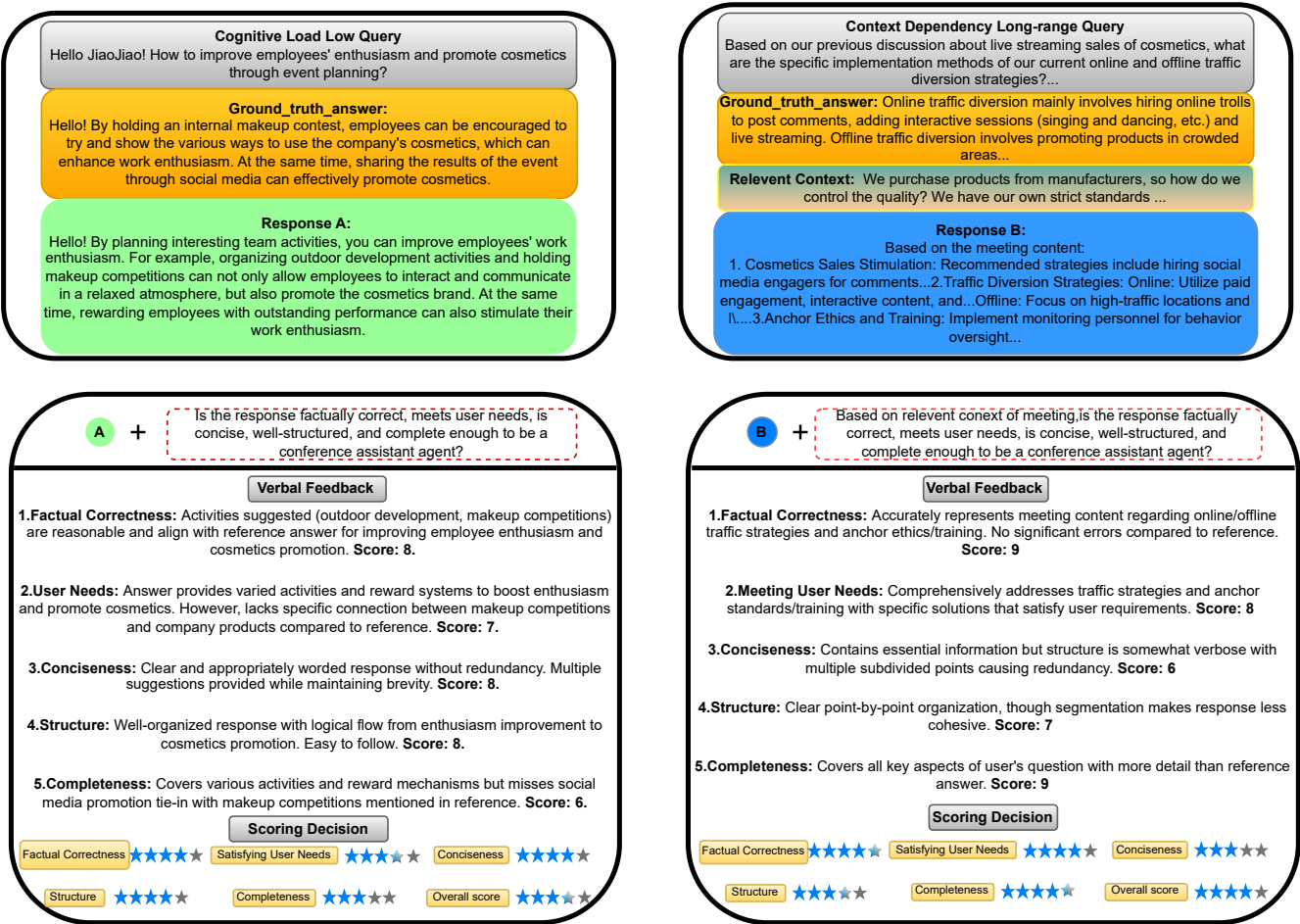


Figure 6: Example analysis of MeetMaster's response evaluation methodology for queries of different types and complexities. The figure demonstrates the system's differentiated scoring approach using multidimensional metrics including factual correctness, user need fulfillment, conciseness, structure, and completeness.

At the end of the evaluation, provide your scoring results in a dictionary format with the following keys:

{'Fact Accuracy': X, 'User Needs Fulfillment': X, 'Conciseness': X, 'Structure': X, 'Completeness': X, 'Overall Score': X}

Remember to evaluate and explain each dimension and ensure that each score is an integer between 1 and 10.

Following examples show the evaluation process of CompassJuder:

User's Question: How can innovation improve the competitiveness of the real estate and renovation industries?

Reference Answer: Innovation can enhance the competitiveness of the real estate and renovation industries by utilizing technology such as big data and AI to optimize market analysis and customer demand prediction, developing eco-friendly materials to reduce environmental impact, implementing smart home technologies to improve quality of life, and offering personalized design services to meet customer needs.

AI Assistant's Answer: Hello, Jiaojiao.

The real estate and renovation industries can innovate in several ways to improve their competitiveness:

1. Use of digital technology: Utilize technologies such as the internet, IoT, and big data to improve business efficiency, provide better customer experience, and reduce operational costs.
2. Promote green and eco-friendly concepts: With growing consumer awareness of environmental protection, the real estate and renovation industries can promote eco-friendly concepts and provide products and services that align with consumer demands.
3. Develop smart homes: Smart homes are a future trend. The real estate and renovation industries can preemptively develop related products and services to capture market opportunities.
4. Improve design quality: By collaborating with designers, the industry can enhance product design to offer more personalized and differentiated products and services.

5. Build brand reputation: Through offering high-quality products and services, businesses can build a strong brand reputation and attract more consumers.

These are some potential innovative strategies, and their implementation should be based on the specific situation of each company.

Evaluation Results:

{'Fact Accuracy': 9, 'User Needs Fulfillment': 8, 'Conciseness': 8, 'Structure': 9, 'Completeness': 8, 'Overall Score': 8}

E.2 Query 2: How should real estate companies adjust their strategic layout and break the homogeneous competition to adapt to market demand and urbanization development trends, while predicting and responding to the risks of a real estate bubble?

Reference Answer: Real estate companies should optimize their commercial real estate model, innovate business strategies, and break free from homogeneous competition. Simultaneously, market research should be strengthened to precisely target consumers and offer diversified products. Regarding the risk of a real estate bubble, companies should adopt measures like reasonable pricing and accelerating capital turnover to avoid crises.

AI Assistant's Answer: Real estate companies should readjust their strategic planning, optimizing existing commercial real estate models to align with market demand and urbanization trends. To counter homogeneous competition, companies need to innovate their business models. Meanwhile, attention should be paid to the risks of a real estate bubble by taking measures such as reasonable pricing and accelerating capital turnover to mitigate potential crises. In addition, companies should strengthen market research, precisely target their customer base, and offer diversified products to meet various consumer demands.

Evaluation Results:

{'Fact Accuracy': 9, 'User Needs Fulfillment': 8, 'Conciseness': 8, 'Structure': 9, 'Completeness': 8, 'Overall Score': 8}

The response provided by the assistant successfully addresses the question by offering practical steps and techniques for enhancing sales and predicting market trends.

E.3 Prometheus Evaluation of AI Assistant's Responses

E.3.1 Query 1: How can innovation improve the competitiveness of the real estate and renovation industries? **Instruction:** How can innovation improve the competitiveness of the real estate and renovation industries?

Reference Answer: Innovation can enhance the competitiveness of the real estate and renovation industries by utilizing technology such as big data and AI to optimize market analysis and customer demand prediction, developing eco-friendly materials to reduce environmental impact, implementing smart home technologies to improve quality of life, and offering personalized design services to meet customer needs.

AI Assistant's Answer: Hello, Jiaojiao.

The real estate and renovation industries can innovate in several ways to improve their competitiveness:

1. Use of digital technology: Utilize technologies such as the internet, IoT, and big data to improve business efficiency, provide better customer experience, and reduce operational costs.

2. Promote green and eco-friendly concepts: With growing consumer awareness of environmental protection, the real estate and renovation industries can promote eco-friendly concepts and provide products and services that align with consumer demands.

3. Develop smart homes: Smart homes are a future trend. The real estate and renovation industries can preemptively develop related products and services to capture market opportunities.

4. Improve design quality: By collaborating with designers, the industry can enhance product design to offer more personalized and differentiated products and services.

5. Build brand reputation: Through offering high-quality products and services, businesses can build a strong brand reputation and attract more consumers.

These are some potential innovative strategies, and their implementation should be based on the specific situation of each company.

Feedback: The response provides a comprehensive approach to enhancing the competitiveness of the real estate and interior design industry. It effectively addresses the user's question by proposing various strategies, such as digitalization, green initiatives, smart home technology, design improvement, and brand reputation enhancement. Each point made is relevant and aligns well with the user's inquiry about innovative methods to improve market position. Furthermore, the structure of the response is clear, making it easy to follow and comprehend. The response demonstrates a good grasp of the topic, with no significant omissions. It offers practical solutions without unnecessary elaboration, adhering closely to the user's needs. In conclusion, the response is informative, relevant, and coherent, making it a solid answer to the user's question.

Score: 5

E.3.2 Query 2: How should real estate companies adjust their strategic layout and break the homogeneous competition to adapt to market demand and urbanization development trends, while predicting and responding to the risks of a real estate bubble? **Instruction:** Based on our previous discussions, how should real estate companies adjust their strategic layout to break homogeneous competition, adapt to market demand and urbanization trends, and anticipate and respond to the risks of a real estate bubble?

Reference Answer: Real estate companies should optimize their commercial real estate model, innovate business strategies, and break free from homogeneous competition. Simultaneously, market research should be strengthened to precisely target consumers and offer diversified products. Regarding the risk of a real estate bubble, companies should adopt measures like reasonable pricing and accelerating capital turnover to avoid crises.

AI Assistant's Answer: Real estate companies should readjust their strategic planning, optimizing existing commercial real estate models to align with market demand and urbanization trends. To counter homogeneous competition, companies need to innovate their business models. Meanwhile, attention should be paid to the risks of a real estate bubble by taking measures such as reasonable pricing and accelerating capital turnover to mitigate potential crises. In addition, companies should strengthen market research, precisely target their customer base, and offer diversified products to meet various consumer demands.

Feedback: The response effectively addresses the core issues discussed by focusing on strategic adjustments, innovation, market research, product diversity, and risk management. It offers clear, actionable steps for businesses to adapt to changing market demands and urbanization trends, while also anticipating potential risks from housing bubbles. Furthermore, the response presents these points in a structured and coherent manner, with a logical progression that enhances its clarity and usability. The information is presented in a concise and straightforward manner without repetition, which makes it easy to understand and implement. Lastly, the response covers all key aspects mentioned in the reference answer and demonstrates a comprehensive understanding of the topic at hand. Hence, the response fulfills all criteria set forth in the score rubric, as it exhibits high accuracy, relevance, simplicity, structure, and completeness in addressing the given instruction. So the overall score is 5.

Score: 5

E.4 4 types of query dataset examples

cognitive_example

Example 1 - Cognitive Low Complexity:

Question: "What ingredients were mentioned in the meeting?"

This is a Cognitive (category: cognitive) Low Complexity (complexity: low) question because it only requires direct extraction of explicitly mentioned ingredients from the dialogue.

Example 2 - Cognitive Medium Complexity:

Question: "What different opinions did participants have about ingredient preparation?"

This is a Cognitive (category: cognitive) Medium Complexity (complexity: medium) question because it requires synthesizing and comparing viewpoints from multiple speakers.

Example 3 - Cognitive High Complexity:

Question: "Based on the discussion, why did some participants prefer certain cooking methods?"

This is a Cognitive (category: cognitive) High Complexity (complexity: high) question because it requires analyzing decision-making reasons and involves inference and evaluation.

context_example

Example 1 - Context No Dependency:

Question: "What kitchen tools are available?"

This is a Context-Dependent (category: context) No Context (complexity: no) question because the answer can be obtained from a single statement without requiring surrounding context.

Example 2 - Context Recent Dependency:

Question: "Why did they change their approach to ingredient preparation?"

This is a Context-Dependent (category: context) Recent Context (complexity: recent) question because it requires understanding the immediate preceding discussion turns.

Example 3 - Context Long-Range Dependency:

Question: "How has the menu plan changed since the start of the discussion?"

This is a Context-Dependent (category: context) Long-Range Context (complexity: long) question because it requires tracking information throughout the entire meeting.

domain example

Example 1 - Domain General Knowledge:
Question: "How many servings of food did they prepare?"
This is a Domain Knowledge (category: domain) General Knowledge (complexity: general) question because it requires no specialized knowledge to understand and answer.

Example 2 - Domain Basic Knowledge:
Question: "Why did they choose specific cooking temperatures and times?"
This is a Domain Knowledge (category: domain) Basic Knowledge (complexity: basic) question because it requires fundamental cooking knowledge.

Example 3 - Domain Expert Knowledge:
Question: "From a food safety perspective, what improvements could be made to their ingredient handling?"
This is a Domain Knowledge (category: domain) Expert Knowledge (complexity: expert) question because it requires professional food safety knowledge.

task example

Example 1 - Task Simple Recording:
Question: "List all ingredients that need to be purchased."
This is a Task Execution (category: task) Simple Recording (complexity: simple) question because it only requires listing explicitly mentioned information.

Example 2 - Task Moderate Processing:
Question: "Summarize the preparation steps and timing for each dish."
This is a Task Execution (category: task) Moderate Processing (complexity: moderate) question because it requires organizing and synthesizing information.

Example 3 - Task Complex Action:
Question: "Based on the discussion, propose improvements to the current cooking process."
This is a Task Execution (category: task) Complex Action (complexity: complex) question because it requires analysis and strategic planning.

E.5 Prompts

We provide the prompts that we used to accomplish problem type routing, query the MeetMaster , create dataset and ground truth.

E.5.1 Planner query classifying prompt. query classifying prompts

As a professional meeting assistant, please analyze the following question’s type and complexity level.

PROBLEM

Who was asked to take over potato duty?

###

Please classify strictly according to these standards:

1. Question Category (category) must be one of:

- cognitive: Questions requiring understanding and processing information
- context: Questions requiring consideration of dialogue context
- domain: Questions requiring specific domain knowledge
- task: Questions requiring task execution

2. Complexity Level (complexity) must be one of the following based on category:

For Cognitive questions:

- low: Simple recall/fact retrieval
- medium: Information synthesis/comparison
- high: Analysis/evaluation/inference

For Context-Dependent questions:

- no: No context dependency
- recent: Recent context dependency
- long: Long-range context dependency
- cross: Cross-meeting context dependency

For Domain Knowledge questions:

- general: General knowledge
- basic: Basic domain knowledge
- expert: Expert domain knowledge

For Task Execution questions:

- simple: Simple recording/listing
- moderate: Information organization/synthesis
- complex: Analysis/planning/strategy

{cognitive_example}

{context_example}

{domain_example}

{task_example}

If type is **Cognitive Load Low** or **Context Dependency None** or Domain Knowledge General, first output "T!". Please provide your response EXACTLY in the following JSON format, ensuring category and complexity values match the above list exactly: {{ "category": "cognitive",

"complexity": "low",

"explanation": "This is a cognitive low complexity question because..." }}

MeetMaster example output

{ "category": "context",

"complexity": "recent",

"explanation": "This is a context-dependent question with recent context dependency because it requires knowledge of who was assigned the task of handling potatoes, which would have been mentioned in the immediate discussion or meeting." }

E.5.2 Planner three-hop reasoning loop prompt. **PLANNING_PROMPT**

```
<PLANNING_PROMPT>
  You are a conference assistant expert. The problem you need to solve is as follows. You can use the following tools to better solve the
  problem. You need to select a tool and output the input of the tool. Output according to the template.
  ### PROBLEM ###
  Who was asked to take over potato duty?
  ###
  ### System tools available ###
  • RagRetrieve – company KB / meeting minutes
  • WebSearch – open-web search
  • MeetSum – abstractive segment summary ###
  ### INPUT: ###
  {
    "keywords": ["k1", "k2", ...],
  }
  ###
  For each keyword:
  – choose the *most appropriate* tool
  – fetch the snippet (≤ 60 words)
  – record its source (url | “Meeting-MM/DD” | “KB-DocID”)
  ### TEMPLATE ###
  Return a JSON list (≤ 6 items):
  [
    "tool": "RagRetrieve", "keywords": [LLM, Agent, ...],
  ] ###
</PLANNING_PROMPT>
```

MeetMaster example output

```
{
  "tool": "RagRetrieve",
  "keywords": ["potato duty", "take over"]
}
```

COMPOSE AND ACT PROMP

<COMPOSE_AND_ACT_PROMPT>

You are a conference assistant expert.You are MeetMaster-Planner composing the final answer.The problem you need to solve is as follows.Your output according to the template

INPUT:

{ "query": "USER_QUERY", "evidence": [tool, source, snippet, ...] # from Planning stage}

###

TASK

1 Write a concise, well-structured answer *in the user's language*.

2 Inline-cite evidence snippets as [1], [2] ... (max 3).

3 If appropriate, select an Action from the catalogue

– ScheduleMeeting, GenerateSlide, SendEmail, None

and fill its minimal JSON input.

###

OUTPUT (strict JSON, no extra text)

{ "answer": "<final answer with [x] citations>", }

###

</COMPOSE_AND_ACT_PROMPT>

MeetMaster example output

{ "answer": "According to the result of RAG, P51 was asked to take over potato duty." }

E.5.3 *Creating Queries Dataset Prompt. Cognitive Complexity*

I'm creating a dataset for meeting assistant agents that can answer questions of varying cognitive complexity. Given the following meeting transcript, please generate three questions (and their answers) that represent different cognitive complexity levels:

1. Low complexity: Simple recall/fact retrieval questions about explicitly stated information
2. Medium complexity: Questions requiring information synthesis or comparison across different parts of the discussion
3. High complexity: Questions involving analysis, evaluation, or inference beyond what is explicitly stated

MEETING TRANSCRIPT:

context

IMPORTANT: Please provide your response EXACTLY in the following format, with no additional text:

Low complexity example: Question: "What was the increase in social media engagement mentioned by P02?"

Ground truth answer: "P02 mentioned a 12% increase in social media engagement after implementing the new content calendar."

The complexity level: "Low complexity"

Explanation of why it belongs to that complexity level: "This question is classified as low complexity because it requires simple recall of an explicitly stated fact from the transcript. The answer can be directly retrieved from P02's statement without requiring any synthesis, comparison, analysis, or inference. It only asks for a specific numerical value (12% increase) that was clearly mentioned in the meeting."

Medium complexity example:

Question: "Could you compare the conversion rates between testimonial posts and overall social media engagement?"

Ground truth answer: "Testimonial posts had a conversion rate of 4.2%, which is higher than the overall social media conversion rate that increased from 2.3% to 3.1%."

The complexity level: "Medium complexity"

Explanation of why it belongs to that complexity level: "This question is classified as medium complexity because it requires synthesizing and comparing information from different parts of the discussion. The answer cannot be found in a single statement but requires gathering data about testimonial post conversion rates and overall social media conversion rates from different parts of the transcript, then making a comparison between these values. This involves connecting related pieces of information that were mentioned separately."

High complexity example: Question: "Based on the discussion, which marketing channel appears to be performing better in terms of customer conversion, and what factors might explain this difference?"

Ground truth answer: "Social media appears to be performing better for conversions, with testimonial posts reaching 4.2% conversion compared to email's 4.8% click-through rate (which is an earlier stage in the conversion funnel). This difference might be explained by the authenticity of user testimonials creating higher trust, the visual nature of social media content, or potentially different audience demographics between channels."

The complexity level: "High complexity"

Explanation of why it belongs to that complexity level: "This question is classified as high complexity because it requires analysis, evaluation, and inference beyond what is explicitly stated in the transcript. Answering requires not only comparing conversion metrics across different marketing channels, but also evaluating which channel is truly performing better (considering that click-through rates and conversion rates represent different stages of the funnel). Furthermore, it asks for possible explanatory factors that might not be explicitly mentioned in the discussion, requiring domain knowledge and inference about marketing principles to propose plausible reasons for the performance differences."

For each question, provide:

- The question
- The ground truth answer
- The complexity level
- Explanation of why it belongs to that complexity level

Context Dependency

I'm creating a dataset for meeting assistant agents that can handle questions with varying degrees of context dependency. Given the following meeting transcript, please generate four questions (and their answers) that represent different context dependency levels:

1. No context: Questions answerable without relying on surrounding context
2. Recent context: Questions requiring information from the immediately preceding discussion
3. Long-range context: Questions requiring information from earlier parts of the meeting
4. Cross-meeting context: Questions requiring information that connects to previously mentioned projects or future work

MEETING TRANSCRIPT:

{context}

IMPORTANT: Please provide your response EXACTLY in the following format, with no additional text:

No context example:

Question: "What are the three database options being discussed?"

Ground truth answer: "The three database options being discussed are an optimized MySQL configuration, migrating to PostgreSQL, or adopting MongoDB as a NoSQL solution."

The context dependency level: "No context"

Explanation of why it belongs to that context dependency level: "This question can be answered directly from the explicit information provided in the transcript without requiring knowledge of the discussion flow, timing, or relationships between statements. The three database options are clearly listed as standalone facts that do not depend on surrounding context to understand."

Recent context example:

Question: "What was the concern P07 raised about the database options?"

Ground truth answer: "P07 raised concerns about the long-term maintenance costs of the different database options."

The context dependency level: "Recent context"

Explanation of why it belongs to that context dependency level: "This question requires understanding the immediately preceding statements in the discussion. To identify P07's specific concern, one needs to follow the recent flow of conversation and identify what P07 just mentioned. The concern is temporally localized to the most recent part of the discussion rather than requiring information from earlier parts or across different discussions."

Long-range context example:

Question: "How does the MongoDB option compare to the budget mentioned earlier in the meeting?"

Ground truth answer: "The MongoDB option would cost around \$60,000, which exceeds the \$50,000 budget allocated for database improvements that was mentioned earlier in the meeting. P01 noted they might need to phase the implementation if they choose this option."

The context dependency level: "Long-range context"

Explanation of why it belongs to that context dependency level: "This question requires connecting information from different, non-adjacent parts of the transcript. To answer it correctly, one must recall the budget amount mentioned much earlier in the meeting (\$50,000) and compare it with the cost of the MongoDB option (\$60,000) discussed later. This requires keeping track of information across the entire timeline of the meeting, not just recent exchanges."

Cross-meeting context example:

Question: "How does the current database discussion relate to the analytics dashboard project mentioned at the end?"

Ground truth answer: "Whatever database solution they choose needs to integrate with the analytics dashboard project that P08's team is developing, as mentioned by P03 at the end of the discussion."

The context dependency level: "Cross-meeting context"

Explanation of why it belongs to that context dependency level: "This question requires connecting the current topic (database selection) with a separate project (analytics dashboard) that bridges to future work. It involves understanding how different projects or initiatives relate to each other across time boundaries, not just within the current meeting. The answer involves recognizing implications for future integration work and dependencies between projects that span beyond the immediate discussion."

For each question, provide:

- The question
- The context dependency level
- The ground truth answer
- Explanation of why it belongs to that context dependency level

Domain Knowledge

I'm creating a dataset for meeting assistant agents that can handle questions requiring different levels of domain expertise. Given the following specialized meeting transcript, please generate three questions (and their answers) that represent different knowledge complexity levels:

1. General knowledge: Questions answerable with basic understanding, requiring no specialized knowledge
2. Domain basics: Questions requiring fundamental understanding of the field's concepts and terminology
3. Domain expertise: Questions requiring deep specialized knowledge, ability to interpret technical data, or make field-specific inferences

MEETING TRANSCRIPT:

{context}

IMPORTANT: Please provide your response EXACTLY in the following format, with no additional text:

General knowledge example:

Question: "How many patients were involved in the clinical trial described in the meeting?"

Ground truth answer: "347 patients with treatment-resistant depression were enrolled in the Phase 2b study of compound XR-27."

The knowledge complexity level: "General knowledge"

Explanation of why it belongs to that knowledge complexity level: "This question requires no specialized knowledge to understand or answer. Anyone who can comprehend basic English and numbers can identify the patient count mentioned in the transcript. No medical or clinical trial expertise is needed to recognize or interpret this straightforward factual information. The question simply asks for a numerical value that is explicitly stated."

Domain basics example:

Question: "What primary measure was used to assess efficacy in this clinical trial, and what was the result for the high-dose group?"

Ground truth answer: "The MADRS (Montgomery-Åsberg Depression Rating Scale) was used to assess efficacy. The high-dose XR-27 group showed a 4.3 point greater reduction in MADRS scores compared to placebo, which was statistically significant with a p-value of 0.008."

The knowledge complexity level: "Domain basics"

Explanation of why it belongs to that knowledge complexity level: "This question requires fundamental understanding of clinical research concepts and terminology. To answer correctly, one needs to know what 'efficacy' means in a clinical context, understand what a rating scale is and why it's used to measure outcomes, and grasp basic statistical concepts like 'reduction in scores,' 'compared to placebo,' and 'p-value' indicating statistical significance. This represents entry-level domain knowledge that someone with basic training in clinical research would possess." Which domain knowledge it belongs to: "Clinical research and psychiatry"

Domain expertise example:

Question: "Based on the QTc prolongation data mentioned and the pharmacokinetic profile, what considerations should be made regarding cardiac monitoring for patients if XR-27 advances to Phase 3 trials?"

Ground truth answer: "Given the 2% incidence of QTc prolongation in high-dose patients (though none exceeded the 500ms threshold), combined with the 18-hour half-life supporting once-daily dosing, Phase 3 trials should include regular ECG monitoring, particularly during dose escalation. Patients with pre-existing cardiac conditions or who are taking other medications known to prolong QTc should be either excluded or monitored more frequently. The lack of significant CYP enzyme interactions suggests lower risk of unexpected QTc effects due to drug-drug interactions, but this should still be verified with dedicated cardiac safety studies."

The knowledge complexity level: "Domain expertise"

Explanation of why it belongs to that knowledge complexity level: "This question requires sophisticated specialized knowledge spanning multiple areas of pharmaceutical research and clinical cardiology. To answer correctly, one must understand cardiac electrophysiology (QTc intervals and their clinical significance), pharmacokinetics (drug half-life and its implications for dosing), drug-drug interactions (particularly CYP enzyme pathways), clinical trial design for cardiac safety monitoring, and regulatory considerations for cardiac risk assessment. The answer requires integrating complex technical information to make expert-level recommendations that would typically come from experienced clinical pharmacologists, cardiologists, or senior clinical research scientists." Which domain knowledge it belongs to: "Clinical pharmacology, cardiology, and pharmaceutical development"

For each question, provide:

- The question
- The knowledge complexity level
- The ground truth answer
- which domain knowledge it belongs to
- Explanation of why it belongs to that knowledge complexity level

Task Execution

I'm creating a dataset for meeting assistant agents that can handle tasks of varying execution complexity. Given the following meeting transcript, please generate three questions/requests (and their responses) that represent different task execution complexity levels:

1. Simple recording: Tasks requiring simple information retrieval or documentation
2. Moderate processing: Tasks requiring synthesis, organization, or summarization of multiple information points
3. Complex action: Tasks requiring analysis, recommendation, strategic planning, or decision support based on the meeting content

MEETING TRANSCRIPT:

{context}

IMPORTANT: Please provide your response EXACTLY in the following format, with no additional text:

Simple recording example:

Question: "Can you list all the team members present in this meeting and their roles?"

Ground truth answer: "The team members present are: P01 (Project Manager), P02 (IT Implementation Lead), P03 (HR Director), and P04 (Finance Manager)."

The task execution complexity level: "Simple recording"

Explanation of why it belongs to that task complexity level: "This task involves straightforward information retrieval without requiring any processing or transformation of the information. The assistant simply needs to identify and list explicitly stated facts from the transcript - the team members' names and their roles. No synthesis, analysis, or inference is needed; the information can be directly extracted and presented in the same form it appears in the meeting."

Moderate processing example:

Question: "Could you create a summary of the project timeline with all the adjusted dates mentioned in the discussion?"

Ground truth answer: "Project Timeline:

Requirements gathering: January (sign-off by January 31st)

System configuration: February to March (completion by March 23rd, reduced from 10 to 8 weeks)

Testing: April to mid-May (extended from 4 to 6 weeks, completion by May 15th)

Training: Rest of May (completion by May 31st, with potential challenges due to P03's annual performance review cycle)

Go-live: June 15th

The entire implementation period spans 6 months."

The task execution complexity level: "Moderate processing"

Explanation of why it belongs to that task complexity level: "This task requires synthesizing and organizing information from multiple parts of the discussion into a coherent structure. The assistant needs to identify all timeline-related information scattered throughout the transcript, understand how dates were adjusted from original plans, organize this information chronologically, and present it in a structured format. This involves more than simple retrieval - it requires comprehension of how different timeline elements relate to each other and the creation of a new information structure (the timeline summary) that wasn't explicitly provided in the original discussion."

Complex action example:

Question: "Based on the concerns raised in the meeting, what are the top three risk factors for this project, and what mitigation strategies would you recommend?"

Ground truth answer: "Top three risk factors and mitigation strategies:

Timeline pressure: The testing phase was initially underestimated, and HR has conflicts during the training period. Mitigation: The team has already extended testing from 4 to 6 weeks, but should also consider creating a detailed testing plan early and potentially offering flexible training schedules to accommodate HR's performance review cycle.

Budget management: The Finance Manager highlighted previous budget overruns (20% on the ERP module). Mitigation: While there is currently a buffer (\$420,000 budget vs. \$395,000 proposal plus 15% contingency), implementing strong change management procedures and regular budget reviews would help prevent scope creep.

Integration challenges: The Finance Manager mentioned the need for proper integration with accounting systems. Mitigation: Schedule dedicated integration testing sessions with the Finance team, document all integration requirements thoroughly during the January requirements phase, and possibly involve integration specialists from the vendor."

The task execution complexity level: "Complex action"

Explanation of why it belongs to that task complexity level: "This task demands high-level analytical thinking and strategic recommendation development based on the meeting content. The assistant must: 1) analyze the entire discussion to identify implicit and explicit risk factors, 2) evaluate and prioritize these risks to determine the top three, 3) formulate specific mitigation strategies that aren't directly mentioned in the transcript, and 4) provide actionable recommendations that integrate project management best practices with the specific context of this project. This requires going beyond the literal content of the meeting to provide decision support that combines domain expertise with critical analysis of the situation."

For each question/request, provide:

- The question/request

- The task execution complexity level
- The ground truth response
- Explanation of why it belongs to that task complexity level